Documentation : Crawler

Karademy Bootcamp Fall 2020

Mentor : Ms. Ashorian

We started the project with the idea of extracting some specific data form an URL link.

*Technologies*

*Languages : python (3.7.8)*

*pip(3.7.8)*

*Libraries :*

1. *urllib : to send request and save the html data*

*installation : pip install urllib*

1. *scrapy : to send request and save the html data and parse it.*

*Installation : pip install scrapy*

1. *bs4 : to send request and save the html data and parse it.*

*Installation : pip install bs4*

1. *re : to find an specific substring in a string variable*
2. *pandas : to save the data as a csv format*

*installation: pip install pandas*

**functions** :

1. requester: it send request to the server and save the data
2. keyword\_finder: it gets the html (which is returned by the requester function ) and a substring and to find it in the text.
3. Link\_finder : it get th e URL and save it into a csv file .

Bugs: at fist the csv method was used to save it as a csv file.

The code :

    #link includes the saved links

    With open (‘file name.csv’, ‘w’) as file:

        csv\_writer=csv.writer(new file ,file name, delimiter= ‘,’)

        csv\_writer.writerow(links)

*The result file, nevertheless, was thousands of lines including only one alphabet like this*

     a

        t

        p

        h

        t

        m

Then the pandas library was applied and it was easier to use and the problem was addressed.

1. tag\_finder : Search through the html file and save all the h tags such as <h1> .

Problems and bugs : Since we are not aware of the numbers of the tags we first check it first and then save the data .

Error message : 1.urllib.error.URLError: <urlopen error [Errno 11002] getaddrinfo failed>

2.'charmap' codec can't encode characters in position 1047-1053: character maps to <undefined>

Both these errors happened whenever the functions tried to open the saved

File or tried to save the data. It seems either ways you are supposed to put the

Encoding (‘utf-8’) when you want to read or write a file.

1. Image\_saver : Search the given link and save all the images

The reddit : it fetch all the picture in the given link and then save them in a html file.

Sources :

1. <https://scrapy.org/>
2. <https://www.youtube.com/watch?v=RkaHdOje-cI&ab_channel=Kite>
3. <https://medium.com/swlh/how-to-use-scrapy-items-05-python-scrapy-tutorial-for-beginners-f25ff2dceaa9>
4. <https://www.youtube.com/watch?v=5o9lucMaQLc&ab_channel=edureka%21>